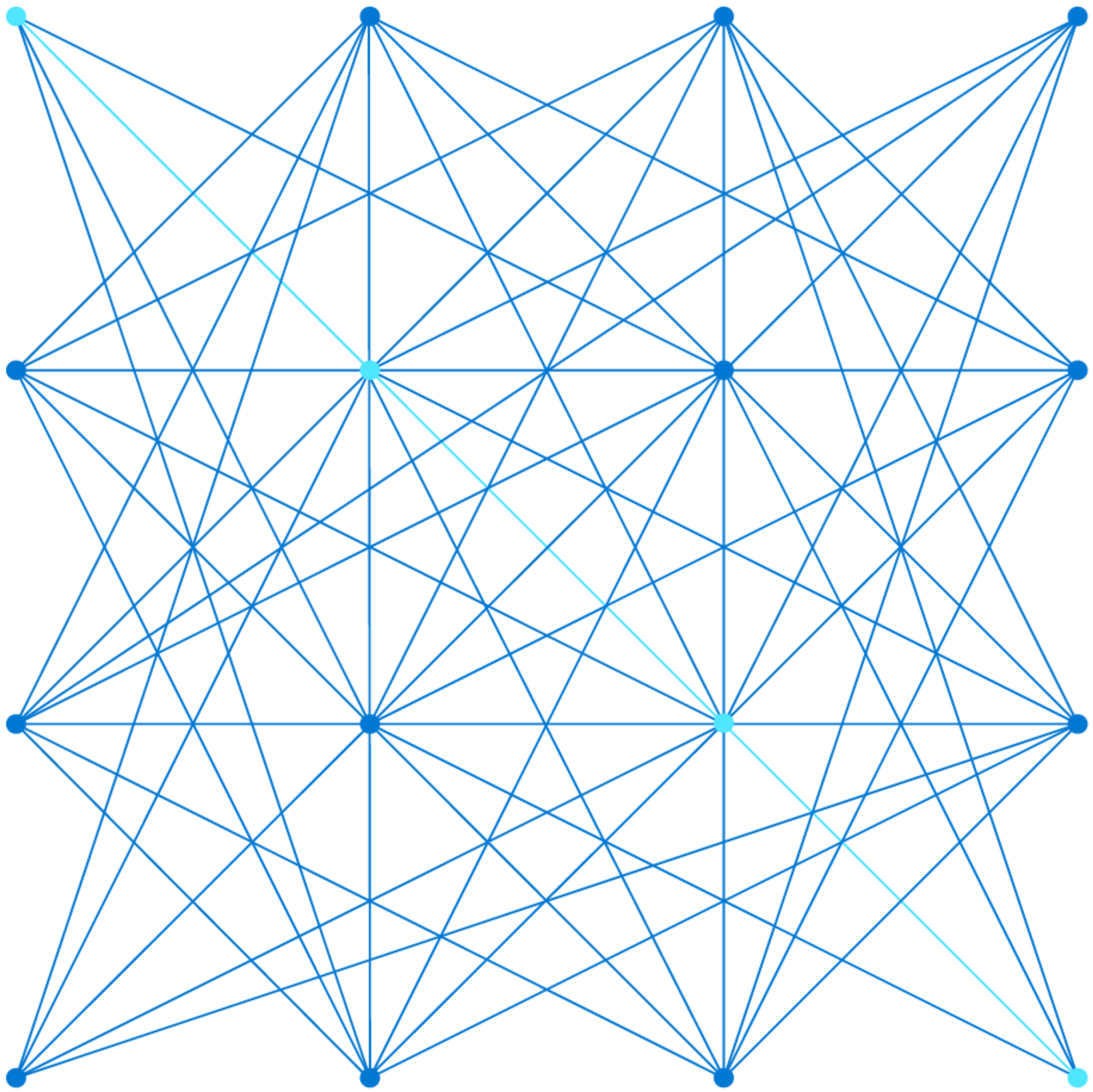
Microsoft Azure

Module 1:
Explore fundamentals of data



Agenda

- Core data concepts
- Data roles and services

Lesson 1: Core data concepts

What is data?

Values used to record information – often representing *entities* that have one or more *attributes*

Structured

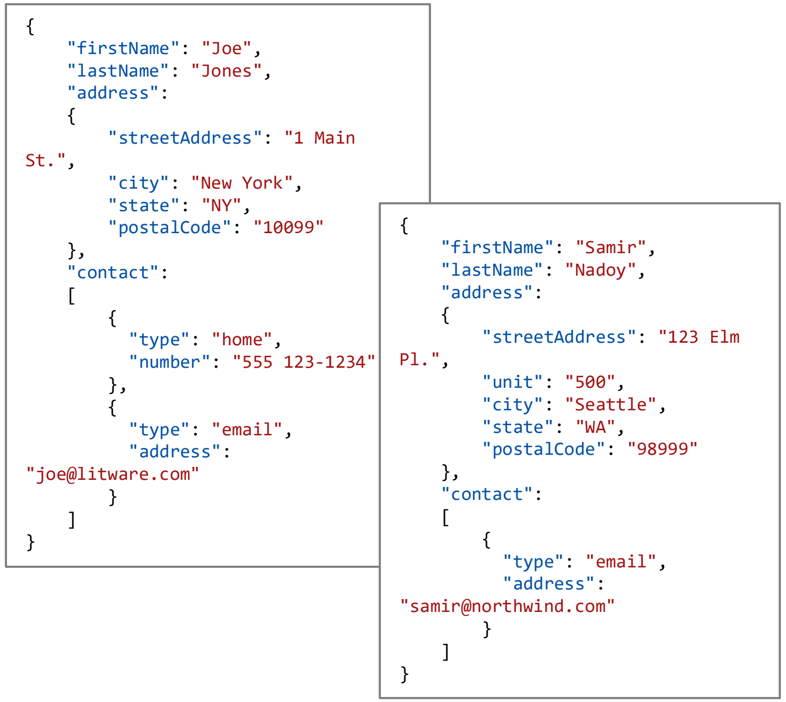
Customer

ID	FirstName	LastName	Email	Address
1	Joe	Jones	joe@litware.com	1 Main St.
2	Samir	Nadoy	samir@northwind.com	123 Elm Pl.

Product

ID	Name	Price
123	Hammer	2.99
162	Screwdriver	3.49
201	Wrench	4.25

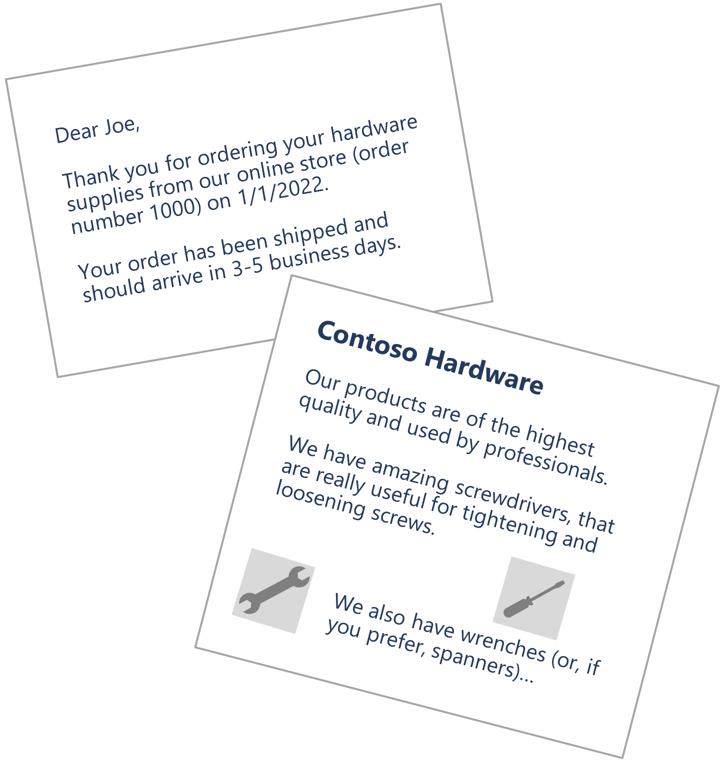
Semi-structured



```
{
    "firstName": "Joe",
    "lastName": "Jones",
    "address":
    {
        "streetAddress": "1 Main
St.",
        "city": "New York",
        "state": "NY",
        "postalCode": "10099"
    },
    "contact":
    [
        {
          "type": "home",
          "number": "555 123-1234"
        },
        {
          "type": "email",
          "address":
"joe@litware.com"
        }
    ]
}
```

```
{
    "firstName": "Samir",
    "lastName": "Nadoy",
    "address":
    {
        "streetAddress": "123 Elm
Pl.",
        "unit": "500",
        "city": "Seattle",
        "state": "WA",
        "postalCode": "98999"
    },
    "contact":
    [
        {
          "type": "email",
          "address":
"samir@northwind.com"
        }
    ]
}
```

Unstructured



Dear Joe,

Thank you for ordering your hardware supplies from our online store (order number 1000) on 1/1/2022.

Your order has been shipped and should arrive in 3-5 business days.

Contoso Hardware

Our products are of the highest quality and used by professionals.

We have amazing screwdrivers, that are really useful for tightening and loosening screws.

We also have wrenches (or, if you prefer, spanners)...

How is data stored?

Files

Delimited Text

```
FirstName,LastName,Email
Joe,Jones,joe@litware.com
Samir,Nadoy,samir@northwind.com
```

JavaScript Object Notation (JSON)

```
{
  "customers":
  [
    { "firstName": "Joe", "lastName": "Jones"},
    { "firstName": "Samir", "lastName": "Nadoy"}
  ]
}
```

Extensible Markup Language (XML)

```
<Customer firstName="Joe" lastName="Jones"/>
```

Binary Large Object (BLOB)

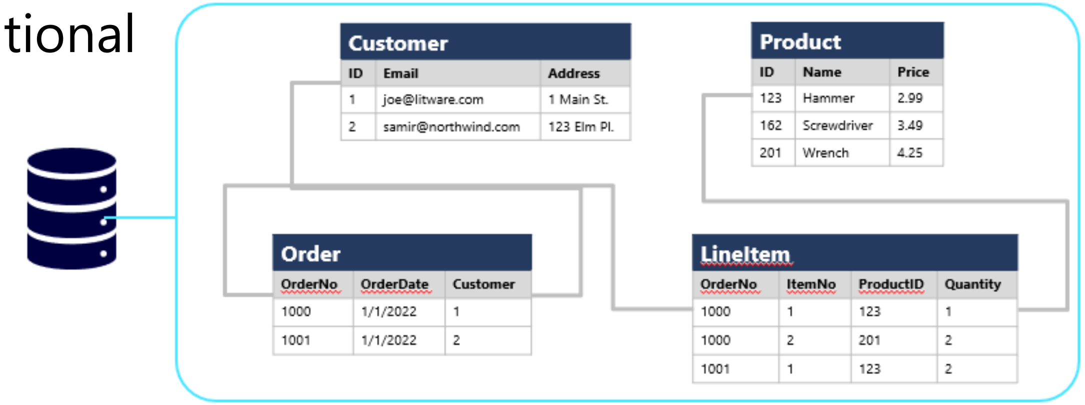
```
10110101101010110010...
```

Optimized formats:

- Avro, ORC, Parquet

Databases

Relational



Customer

ID	Email	Address
1	joe@litware.com	1 Main St.
2	samir@northwind.com	123 Elm Pl.

Product

ID	Name	Price
123	Hammer	2.99
162	Screwdriver	3.49
201	Wrench	4.25

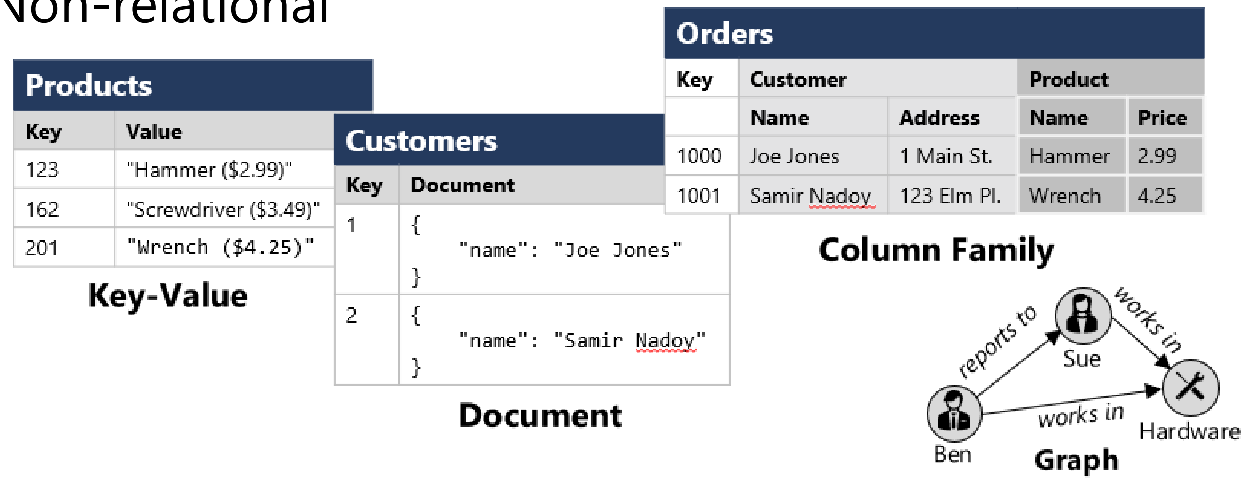
Order

OrderNo	OrderDate	Customer
1000	1/1/2022	1
1001	1/1/2022	2

LineItem

OrderNo	ItemNo	ProductID	Quantity
1000	1	123	1
1000	2	201	2
1001	1	123	2

Non-relational



Products

Key	Value
123	"Hammer ($2.99)"
162	"Screwdriver ($3.49)"
201	"Wrench ($4.25)"

Key-Value

Customers

Key	Document

1	{
	"name": "Joe Jones"
	}

2	{
	"name": "Samir Nadoy"
	}

Document

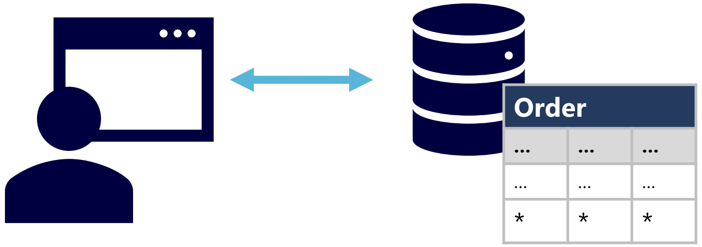
Orders

Key	Customer		Product	
	Name	Address	Name	Price
1000	Joe Jones	1 Main St.	Hammer	2.99
1001	Samir Nadoy	123 Elm Pl.	Wrench	4.25

Column Family

Graph

Transactional data workloads



Data is stored in a database that is optimized for *online transactional processing* (OLTP) operations that support applications

A mix of *read* and *write* activity

For example:

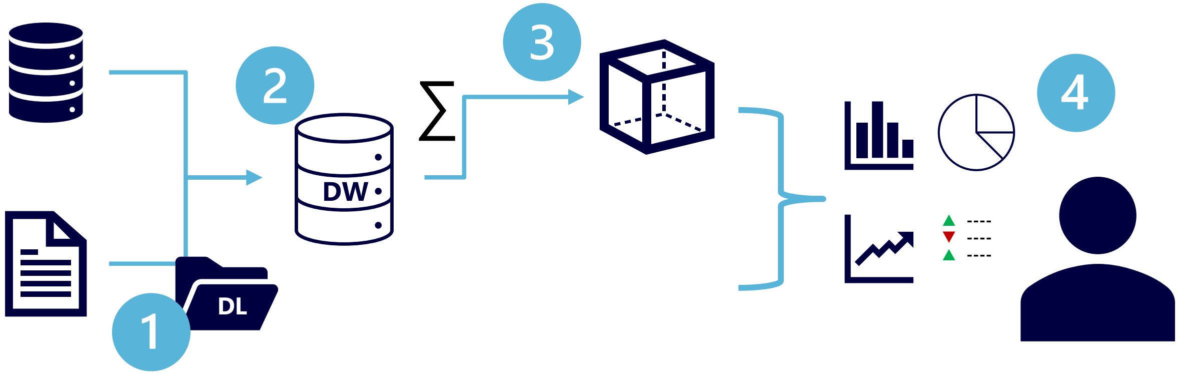
- Read the *Product* table to display a catalog
- Write to the *Order* table to record a purchase

Data is stored using *transactions*

Transactions are "ACID" based:

- **Atomicity** – each transaction is treated as a single unit of work, which succeeds completely or fails completely
- **Consistency** – transactions can only take the data in the database from one valid state to another
- **Isolation** – concurrent transactions cannot interfere with one another
- **Durability** – when a transaction has succeeded, the data changes are persisted in the database

Analytical data workloads



1. Data files may be stored in a central *data lake* for analysis
2. An extract, transform, and load (ETL) process copies data from files and OLTP databases into a *data warehouse* that is optimized for *read* activity
3. Data in the data warehouse may be aggregated and loaded into an online analytical processing (OLAP) model, or *cube*
4. The data in the data lake, data warehouse, and analytical model can be queried to produce reports and dashboards

Lesson 1: Knowledge check

How is data in a relational table organized?

☑ Rows and Columns

☐ Header and Footer

☐ Pages and Paragraphs

Which of the following is an example of unstructured data?

☐ A comma-delimited text file with *EmployeeID*, *EmployeeName*, and *EmployeeDesignation* fields

☑ Audio and Video files

☐ A table within relational database

What is a data warehouse?

☐ A non-relational database optimized for read and write operations

☑ A relational database optimized for read operations

☐ A storage location for unstructured data files

Lesson 2: Data roles and services

Data professional roles

Database Administrator

Database provisioning, configuration and management

Database security and user access

Database backups and resiliency

Database performance monitoring and optimization

Data Engineer

Data integration pipelines and ETL processes

Data cleansing and transformation

Analytical data store schemas and data loads

Data Analyst

Analytical modeling

Data reporting and summarization

Data visualization

Microsoft cloud services for data

Data stores

Azure SQL

- Family of SQL Server based relational database services

Azure Database for open-source

- Maria DB, MySQL, PostgreSQL

Azure Cosmos DB

- Highly scalable non-relational database system

Azure Storage

- File, blob, and table storage
- Hierarchical namespace for data lake storage

Data engineering and analytics

Azure Data Factory

- Data pipelines

Azure Synapse Analytics

- Integrated, end-to-end analytics
- Pipelines, SQL, Apache Spark, Data Explorer …

Azure Databricks

- Apache Spark analytics and data processing

Azure HDInsight

- Apache open-source platform

Azure Stream Analytics

- Real-time data processing for IoT solutions

Azure Data Explorer

- Real-time data analysis for logs and telemetry

Microsoft Purview

- Enterprise data governance
- Data mapping and discoverability

Microsoft Power BI

- Analytical data modeling
- Interactive data visualization

others…

Lesson 2: Knowledge check

Which one of the following tasks is the responsibility of a database administrator?

- ☑ Backing up and restoring databases
- □ Creating dashboards and reports
- □ Creating pipelines to process data in a data lake

Which role is most likely to use Azure Data Factory to define a data pipeline for an ETL process?

- □ Database Administrator
- ☑ Data Engineer
- □ Data Analyst

Which single service would you use to implement data pipelines, SQL analytics, and Spark analytics?

- □ Azure SQL Database
- □ Microsoft Power BI
- ☑ Azure Synapse Analytics

Microsoft Azure